Data Engineering Roadmap

- **Programming Languages**
  - *Python (Recommended)*
  - *Java*
  - *Scala*

- **Data Exploration Libraries (Python Only)**
  - *Pandas*
  - NumPy
  - Matplotlib

- **Operating Systems & Scripting**
  - *Linux/Unix Commands*
  - *Shell Scripting*
  - *Cron Jobs*

- **Data Structures & Algorithms (Easy-Medium Level Only)**
  - *Arrays*
  - *Strings*
  - *Linked List*
  - *Stack*
  - *Queue*
  - *Tree (Basics)*
  - *Graph (Basics)*
  - *Dynamic Programming*
  - *Searching*
  - *Sorting*

- **Database Management Systems**
  - *Understand RDBMS and its use cases*
  - *Schema Types*

- *ER Diagram*
- *ACID Properties*
- *Transactions*
- *Concurrency Control*
- *Deadlock*
- *Indexing*
- *Hashing*
- *Normalization Forms*
- *Views*
- *Stored Procedures*

- **SQL**
  - *Basics Of DDL, DML, DCL*
  - *All Types Of Joins*
  - *Subqueries*
  - *Group By*
  - *Case-When Statement*
  - *Common Table Expression (With Clause)*
  - Window Functions
  - Pivoting

- **BigData Terminologies**
  - *What is BigData?*
  - *5 V's of BigData*
  - *Distributed Computation*
  - *Distributed Storage*
  - *Vertical vs Horizontal Scaling*
  - *Commodity Hardware*
  - *Clusters*
  - *File formats*
    a. *CSV*
    b. *JSON*
    c. *AVRO*

    d. *Parquet*
    e. *ORC*
  - *Type of Data*
    a. *Structured*
    b. *Unstructured*
    c. *Semi-structured*

- **Data Warehousing**
  - *OLAP vs OLTP*
  - *Dimension Tables*
  - *Fact Tables*
  - *Star Schema*
  - *Snowflake Schema*
  - *Warehouse Designing Questions*
  - *Slowly Changing Dimensions (SCD)*

- **BigData Frameworks**
  - *Apache Hadoop (Architecture Understanding Most Important)*
    a. *HDFS*
    b. *Map-Reduce (Coding part not needed)*
    c. *Yarn*
  - *Apache Hive*
    - *How to load data in different file formats*
    - *Internal Tables*
    - *External Tables*
    - *Querying table data stored in HDFS*
    - *Partitioning*
    - *Bucketing*
    - *Map-Side Join*
    - *Sorted-Merge Join*
    - *UDFs in Hive*
    - *SerDe in Hive*
  - *Apache Spark (Most Important)*

        - *Spark Core*
        - *Spark SQL*
        - *Spark Streaming*
    - *Apache Flink (Real-Time Data Processing)*
    - *Apache SQOOP*
    - *Apache NIFI*
    - *Apache FLUME*

- **Schedulers/Workflow Managers**
    - *Apache Airflow*
    - *Apache NIFI*
    - *Azkaban*

- **NoSQL Databases**
    - *HBase*
    - *Cassandra*
    - *ElasticSearch*
    - *MongoDB*

- **Messaging Queue**
    - *Apache Kafka*

- **Dash Boarding Tools**
    - *Tableau*
    - *PowerBI*
    - *Grafana*
    - *Kibana*

- **BigData Services in Cloud (AWS)**
    - *On-demand Machines*
        - *AWS EC2*
    - *Access Management*

    - *AWS IAM*
- *For Storing and Accessing Credentials*
    - *AWS Secret Manager*
- *Distributed File Storage*
    - *AWS S3*
- *Transactional Database Services*
    - *AWS RDS*
    - *AWS Athena*
- *Data Warehousing Service*
    - *AWS Redshift*
- *NoSQL Database Services*
    - *AWS Dynamo*
- *Serverless*
    - *AWS Lambda*
- *ETL Services*
    - *AWS Glue*
- *Scheduler*
    - *AWS CloudWatch*
- *Distributed Data Computation*
    - *AWS EMR*
- *Messaging Queue*
    - *AWS SNS*
    - *AWS SQS*
- *Real-Time Data Processing*
    - *AWS Kinesis*